Item Response Theory Analysis of the Recoded Internet Gaming Disorder Scale-9 Short-Form (IGDS9-SF)

Gomez, R, Stavropoulos, Vasileios, Beard, C and Pontes, HM (2019) Item Response Theory Analysis of the Recoded Internet Gaming Disorder Scale-9 Short-Form (IGDS9-SF). International Journal of Mental Health and Addiction, 17 (4). pp. 859-879. ISSN 1557-1874

Int J Ment Health Addiction (2019) 17:859–879
https://doi.org/10.1007/s11469-018-9890-z

ORIGINAL ARTICLE

Item Response Theory Analysis of the Recoded Internet Gaming Disorder Scale-Short-Form (IGDS9-SF)

Rapson Gomez[1] · **Vasileios Stavropoulos**[2,3] · **Charlotte Beard**[4] · **Halley M. Pontes**[5]

Published online: 26 March 2018

Abstract Based on the nine criteria for Internet gaming disorder (IGD) in DSM-5, the Internet Gaming Disorder Scale 9-Short Form (IGDS9-SF; Pontes and Griffiths 2015) is the most widely used questionnaire for assessing IGD. The present study examined support for the unidimensional factor structure of the instrument, with a group of 868 adolescent and adult gamers from the USA, with criteria recoded as present or absent. The two-parameter logistic model (2PLM) was used to examine the item response theory properties of the criteria included in the measure. Confirmatory factor analysis supported the one-factor model. The 2PLM analysis indicated that all the criteria were strong discriminators of high and low latent IGD. Furthermore, the items measured more of the GAD dimension and with more precision from around +2 *SD* from the mean trait level. The implications of the findings for interpreting the IGDS9-SF scores for clinical practice are discussed.

Keywords IGDS9-SF · One-factor model · Item response theory · 2PLM

The latest fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) of the American Psychiatric Association [APA] (American Psychiatric Association 2013) includes Internet gaming disorder (IGD) as a condition worthy of future study. According to the DSM-5, IGD comprises a behavioral pattern encompassing persistent and recurrent use of

✉ Halley M. Pontes
contactme@halleypontes.com

[1] Federation University, University Dr, Mount Helen, Ballarat, VIC 3350, Australia

[2] Faculty of Psychology, School of Philosophy University of Athens Panepistimiopoli, Ilissia, Athens 157 84, Greece

[3] Cairnmillar Institute, 391/393 Tooronga Rd, Hawthorn East, VIC 3123, Australia

[4] Palo Alto University, 1791 Arastradero Rd, Palo Alto 94304 CA, USA

[5] Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ, UK

the Internet to engage in online and offline games, leading to significant impairment or distress over a period of 12 months. The diagnosis of IGD requires the endorsement of at least five of the following nine criteria: preoccupation with Internet games (criterion 1); withdrawal behaviors when Internet gaming is taken away (criterion 2); tolerance, resulting in the need to spend increasing amounts of time engaged in Internet games (criterion 3); unsuccessful attempts to control participation in Internet games (criterion 4); loss of interest in previous hobbies and entertainment as a result of, and with the exception of, Internet games (criterion 5); continued excessive use of Internet games despite knowledge of psychosocial problems (criterion 6); deceiving family members, therapists, or others regarding the amount of Internet gaming (criterion 7); use of Internet games to escape or relieve negative moods (criterion 8); and jeopardizing or losing a significant relationship, job, or education or career opportunity because of participation in Internet games (criterion 9). The DSM-5 has, however, indicated that more empirical evidence is needed to investigate the suitability and clinical utility of these nine diagnostic criteria in case IGD is to be fully recognized as a bona fide addiction that is featured in the next revisions of the DSM. Accordingly, the aim of the current study was to use item response theory (IRT) procedures to examine the psychometric properties of these criteria, as provided in the Internet Gaming Disorder Scale 9-Short Form (IGDS9-SF; Pontes and Griffiths 2015), in a group of adolescent and adult gamers from the United States of America (USA).

Since the publication of DSM-5, a number of self-rating scales comprising the nine IGD criteria have been developed. Of these, as noted by Wu et al. (2017a), the most widely used questionnaire is the IGDS9-SF (Pontes and Griffiths 2015). Based on the criteria for IGD in the DSM-5, the IGDS9-SF was developed to assess the severity of the nine IGD criteria for both online and/or offline detrimental effects of gaming activities occurring over the previous 12-month period. It is suitable for children and adults, and is claimed to be efficient for large-scale surveys (Wu et al. 2017b). In terms of assessment of disordered gaming criteria, the IGDS9-SF and DSM-5 IGD criteria are highly comparable. Each item of the IGDS9-SF is rated on a five-point scale ("never," scored 1; "rarely," scored 2, "sometimes," scored 3; "often," scored 4; and "very often," scored 5). In terms of cutoff diagnostic threshold for the IGDS9-SF, Pontes and Griffiths (2016) (see also Pontes et al. 2016) have suggested that the endorsement of five or more of the nine criteria in the IGDS9-SF, on the basis of answering "very often," could be considered as the basis for clinical diagnosis, thus mirroring the diagnostic framework suggested in the DSM-5 (APA 2013).

In the initial IGDS9-SF scale development and validation study, Pontes and Griffiths (2015) reported that a single IGD factor was supported for a large and heterogeneous group of English-speaking adolescent and adult gamers ($N = 532$) through exploratory factor analysis (EFA) and parallel analysis. The EFA revealed that the single IGD domain explained 45.4% of the total variance, which was supported by confirmatory factor analysis (CFA) in a second sample ($n = 528$). Criterion validity was also supported in both samples, as respondents' IGDS9-SF total scores correlated significantly with their reported weekly gameplay time and the total score obtained in the Internet Gaming Disorder-20 Test (IGD-20; Pontes et al. 2014; Pontes, 2016). The IGD-20 test is an extended measure to assess IGD that is aligned with Griffiths's (2005) theoretical framework of the components model of addiction (i.e., salience, mood modification, tolerance, withdrawal criteria, conflict, and relapse) (Pontes and Griffiths 2014). In relation to reliability, the IGDS9-SF has demonstrated adequate levels of internal consistency (Cronbach's alpha ≥ .80), and acceptable measurement precision in terms of standard error of measurement (SEM), with nonsignificant floor or ceiling effects across items. Taken together, the initial development and psychometric validation of the IGDS9-SF posited it as a sound standardized measure that aligns with the goals of IGD as a condition for further study, with promising potential for both research and clinical use (Pontes and Griffiths 2015).

To date, a number of studies have extended the exploration of the psychometric properties of the IGDS9-SF in different cultural contexts (Monacis et al., 2016; Pontes and Griffiths 2016; Pontes et al. 2016; Stavropoulos et al. 2017; Pontes et al. 2017). More specifically, the psychometric properties of the IGDS9-SF have been extensively investigated across several cross-cultural studies. For example, Monacis et al. (2016) investigated a sample of Italian adolescents and adult gamers from the general community, Pontes and Griffiths (2016) examined Portuguese adolescents from the general community, Pontes et al. (2016) examined Slovenian adolescents from the general community, and Wu et al. (2017a) examined Iranian adolescents from the general community. Stavropoulos et al. (2017) examined adult gamers from the USA, Australia, and the UK, and Pontes et al. (2017) investigated adult gamers from the US, the UK, and India. Overall, the results across these studies have generally confirmed the initial findings reported by Pontes and Griffiths (2015), supporting the scale's one-factor model (based on CFA), criterion, convergent, and discriminant validity in terms of expected significant correlations with relevant external variables, in addition to various forms of reliability (e.g., internal consistency, composite reliability, SEM, average variance extracted, and factor determinacy). Furthermore, in the Italian study of the IGDS9-SF, Monacis et al. (2016) found support for measurement invariance (i.e., configural, metric, and scalar) across sex and age groups. Wu et al. (2017a) found measurement invariance across gender and time spent online gaming per week, and Stavropoulos et al. (2017) and Pontes et al. (2017) reported support for partial measurement invariance across respondents from the USA, Australia, and the UK, and the USA, the UK, and India respectively.

The Item Response Theory Significance on Examining IGDS9-SF

Overall, despite the fact that the IGDS9-SF has only been developed relatively recently, there is a large body of studies supporting its psychometric properties, in particular, its factor structure (based on EFA and CFA); validity (e.g., criterion, convergent, and discriminant); reliability (e.g., internal consistency, composite reliability, SEM, average variance extracted, and factor determinacy); full measurement invariance (e.g., configural, metric, and scalar) across gender, age groups, and time spent online gaming per week; and partial measurement invariance across a number of national groups (Hawi and Samaha 2017; Monacis et al. 2016; Pontes and Griffiths 2016; Stavropoulos et al. 2017; Wu et al., 2017a). However, as will be noticed, the psychometric data described above are all based on classical test theory (CTT). Item response theory (IRT) is another form of psychometric validation that can be used to evaluate the psychometric properties of measures that is distinct in both theory and form from CTT (Embretson and Reise 2000). IRT is a modern measurement theory that is model-based, and it envisages the relationship between responses to an item and the latent trait that the item is intended to measure (Birnbaum 1968; Embretson and Reise 2000). It has been argued that IRT has many advantages over CTT for evaluating the psychometric properties of instruments (Embretson and Reise 2000). First, while in order to obtain a score for a trait, CTT uses the sum of the various items assumed to represent the trait, using IRT allows researchers to obtain a trait score for each item. Second, CTT assumes and provides one reliability (such as internal consistency) value and one standard error of measurement value for all levels of the scores obtained in a measure. In contrast, IRT provides the reliability of each item at different levels of the underlying trait, controlling for the characteristics (e.g., difficulty) of the different items in the scale. Third, because CTT-extracted psychometric properties, such as reliability, item-total correlation, and standard error of measurement are sample dependent, these properties can vary significantly across a range of samples, and thus fail to generalize to a given population. In contrast, IRT psychometric properties are assumed to be sample independent or group invariant within a linear transformation.

In order to assess an item's representation of the latent trait, the IRT approach generates graphs of trace lines or curves for each item (Raykov and Marcoulides 2011). These curves demonstrate the probability of a positive response to the items as a function of the underlying trait. For an item with dichotomous responses (such as criterion "present" or criterion "absent"), the trace lines are called item characteristic curves (ICCs). Two common item parameters associated with the ICC are the item difficulty (or threshold) parameter and the item discrimination parameter. The difficulty parameter (β) indicates the point on the scale of the latent trait where a person has a 0.5 probability of responding positively to the item, while the item discrimination parameter (α) is the ability of an item (determined by the slope of the ICC) to discriminate people with different levels of the underlying trait below and above the threshold parameter (Steinberg and Thissen 1995). The graphs of the top panel of Fig. 1 display ICCs for two hypothetical items. Compared to the right graph, the left graph shows a higher difficulty parameter (shown as b), thereby indicating that higher trait values are needed for a positive response to the item shown on the left side. The ICC of the graph on the left also shows a higher item discrimination parameter (shown as a), suggesting that this item will be better able to discriminate the underlying trait for those with low and high levels of the underlying trait, compared to the item shown on the graph shown on the right side.

Besides the ICC, α and β, IRT models can also generate item information function (IIF), and test information function (TIF). The IIF indicates the effectiveness or reliability of an item to measure the latent trait at different levels of the trait, while the TIF provides the effectiveness or reliability of the test (i.e., all items together) to measure the latent trait at different levels of the trait (Embretson and Reise, 2000). The bottom panel of Fig. 1 shows examples of an IIF (left graph) and a TIF (right graph). The IIF graph indicates that the precision of the item is especially high around the mean value of the trait. The TIF graph illustrates that the test has high precision around a trait value ranging from 0.5 to 2, with low precision at other levels of the trait. In addition to IIF and TIF information, the test characteristic curve (TCC) can also be used to indicate the trait scores for different numbers of positive responses on the scale.

The Two-Parameter Logistic Model

Currently, there are several IRT models that are implemented based on the types of scores used for the analysis. In this respect, models that are appropriate for dichotomous scores are the one-parameter logistic model (1PLM) and the two-parameter logistic model (2PLM; Birnbaum 1968). The 1PLM, mathematically equivalent to the Rasch model, assumes that all items have the same discrimination parameter value, and provides difficulty (threshold) parameter estimates (β) for items. In contrast, 2PLM provides difficulty (threshold) and discrimination parameter estimates for each of the items. For the 2PLM, the mathematical function for the ICC for the probability of a positive response to an item j for a person with latent trait θ [i.e, $P_j(\theta)$] is as follows:

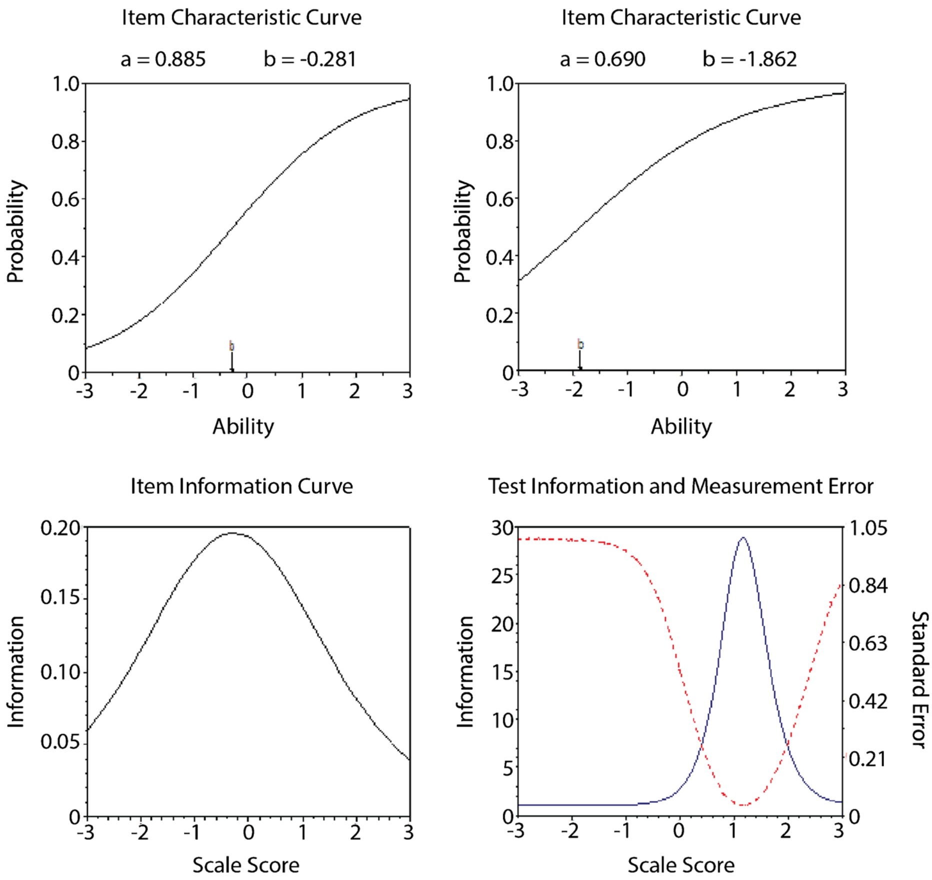

Fig. 1 Graphs used to explain item response theory in the text

$$P_j(\theta) = \frac{\exp\left[\alpha_j\left(\theta - \beta_j\right)\right]}{1 + \exp\left[\alpha_j\left(\theta - \beta_j\right)\right]}$$

As will be noticed in the above equation, in the 2PLM the probability of a positive response for an individual is a function of a person's trait level (θ), and the item discrimination (α_j) and threshold (β_j) parameters. For each item, the following psychometric properties can be computed: item characteristic curve (ICC), discrimination parameter (α), threshold parameter (β), and IIF values. In addition, for the overall scale, the TIF curve can also be produced. At this point, it should be noted that the implementation of the IRT 2PLM model was chosen in the present study for three reasons: (a) to align with DSM 5 and clinical psychology practice diagnostic guidelines, which are based on dichotomous ratings (presence or absence of a behaviour/symptom) paired with functional consequences (American Psychiatric Association, 2013; Keeley et al. 2016); (b) to align and to provide relative comparability with past IRT studies of IGD measures (that also implemented the 2PLM model; Király et al., 2015); and (c) to align with IRT studies of assessment instruments in the broader field of behavioral addictions (see *Problem Gambling Severity Index*; Sharp et al. 2012). In that context, the 2PLM model cut off point for the present study has been placed between "often" and "very often" to correspond (be consistent) with the more stringent binary scoring method recommended by Pontes and Griffiths (2015) and Pontes et al. (2016).

At least two studies have used IRT to examine the properties of the IGD criteria (Király et al., 2015; Wu et al., 2017a). Wu et al. (2017a) applied the Rasch model on the ratings of the IGDS9-SF, and reported wide variability in the item difficulty parameters. They ranged from + 1.57 SD from the mean (for "preoccupation") to − 1.06 SD from the mean (for "negative consequences"). The symptom for "withdrawal" also had high difficulty parameter at + 1.4 SD from the mean, and the values for the other criteria ranged from − 73 to + 26 *SD* from the mean, thereby suggesting that they were more representative of the severity of the criteria around the mean trait level. The study by Király et al. (2015) examined the IRT properties in a large group of Hungarian gamers (age ranging from 14 to 64 years) for the nine IGD criteria provided in the Internet Gaming Disorder Test-10 (IGDT-10; Király et al. 2015). Like the IGDS9-SF, the IGDT-10 comprises the nine criteria proposed by DSM-5 for the diagnosis of IGD. However, criterion 9 in the IGDT-10 is split into two items, and, unlike the IGDS9-SF, each item is rated on a three-point scale ("never," scored 0; "sometimes," scored 1; and "often," scored 1). For the IRT analyses conducted by Király et al. (2015), these options were recoded into "criterion present" (if the "often" option was endorsed) and "criterion not present" (if the "never" or the "sometimes" options were endorsed; scored 0). As items 9 and 10 were related to the same criterion, they were combined in the scoring, so that endorsing "often" on either item 9 or item 10 or both scored only 1 point. The IRT findings, based on the 2PLM for the difficulty parameter, indicated that the nine criteria represented different levels of severity, with "continuation despite problems," "preoccupation," "negative consequences," and "escape" being associated with relatively lower severity of IGD (ranging from + 1.27 to + 1.83 *SD* from the mean), whereas "tolerance," "loss of control," "giving up other activities," "deception," and "withdrawal" associated with relatively more severe levels of IGD (ranging from + 2.29 to + 2.54 *SD* from the mean). The discrimination parameters also indicated that the nine criteria have different abilities to distinguish those with high and low levels of the IGD trait, with "preoccupation," "loss of control," "escape," and "deception" having relatively lower ability (ranging from 0.66 to 0.98); and "withdrawal," "tolerance," "giving up other activities," "continuation despite problems," and "negative consequences" having relatively better ability (ranging from 1.04 to 1.35). Relative to the other criteria, "preoccupation" and "escape" provided very little information (reliability) to IGD severity across the entire trait spectrum.

Apart from the study reported by Wu et al. (2017a), no other study has examined the IRT properties of the IGDS9-SF. In this respect, although the comparability of the items across the IGDT-10 and the IGDS9-SF could imply that the IRT findings reported by Király et al. (2015) may be applicable to the IGDS9-SF, there are important differences across the IGDT-10 and the IGDS9-SF that make this unlikely. Specifically, (1) the IGDT-10 has a three-point scale, whereas the IGDS9-SF has a five-point scale ("never," scored 1; "rarely," scored 2; "sometimes," scored 3; "often," scored 4; and "very often," scored 5) for its items, (2) "very often" is used to infer the presence of criteria in the IGDS9-SF (Pontes and Griffiths 2016; Pontes et al. 2016), whereas in the IGDT-10, endorsement of "often" is used to infer the presence of the criteria (Király et al., 2015). Such differences could potentially influence parameter estimates as, in IRT, responses to items are used to obtain continuous scaled estimates of the underlying trait, called theta (θ). In most computer programs, the trait value is set at a mean of zero and a standard deviation of one. Since the differences in the scale and response categories across the IGDS9-SF and IGDT-10 will lead to different responses by the same individual, the metrics for the IGDS9-SF and IGDT-10 would be different. Consequently, their parameters cannot be directly compared. This means that it cannot be assumed that the

IRT findings found for the IGDT-10 in the Király et al. (2015) study are fully applicable to the IGDS9-SF. Therefore, IRT estimates of the IGDS9-SF need to be directly tested. Given the wide and varied use of the IGDS9-SF, its brevity, its alignment with DSM-5, and existing information on its CTT-based psychometric properties, having additional information on its IRT-based psychometric properties is paramount for better and more accurate application and use of the of this scale. Furthermore, diagnosis in clinical psychology is based on dichotomous ratings paired with functional consequences. Therefore, using the most extreme end of a five-point response scale may offer more information about the reflection of each criterion to the construct of IGD and the severity and weight of each IGD criterion with regard to diagnostic and treatment considerations.

Due to the need for further evaluation as detailed above, the aim of the present study was to examine the IRT properties of the IGD criteria as provided in the IGDS9-SF. To increase the clinical utility of the findings, the criterion rating scores were recoded as binary scores, reflecting either criterion being present or criterion not being present. As mentioned above, this type of dichotomous scoring aligns with the way IGD criteria are coded during clinical interview and diagnosis. The 2PLM was used for the IRT analyses. As the 2PLM requires unidimensionality in the criteria comprising the measure, this was tested prior to the IRT analyses. For each criterion, the following psychometric properties were examined: item characteristic curve (ICC), discrimination parameter (α), difficulty or threshold parameter (β), and item information function (IIF) values. In addition, for the overall scale, the test information function (TIF) curve and test characteristic (TCC) were examined. Although it has been argued that the findings reported in the study by Király et al. (2015) for the IGDT-10 are unlikely to be fully applicable to the IGDS9-SF, in the absence of other IRT data, the findings from the Király et al. (2015) study were utilized to inform the present hypotheses. Specifically, it was hypothesized that while there would be some degree of variability, all the criteria were expected to show high discrimination parameter values. Criterion indicators of "preoccupation," "loss of control," "escape," and "deception" were expected to demonstrate low discrimination values for IGD, while "withdrawal," "tolerance," "giving up other activities," "continuation despite problems," and "negative consequences" were expected to demonstrate strong discrimination. It was additionally hypothesized that the nine criteria would vary in terms of their difficulty parameters. In particular "continuation despite problems," "preoccupation," "negative consequences," and "escape" were speculated to be associated with relatively lower severity of IGD, whereas "tolerance," "loss of control," "giving up other activities," "deception," and "withdrawal" were speculated to be associated with relatively more severe levels of IGD.

Method

Participants

The sample was comprised a total of 868 Internet gamers from the USA, with ages ranging between 16 and 70 years ($m_{\mathrm{age}} = 28.64$; $SD = 8.79$ years). The sample included 345 females (39.75%; $m_{\mathrm{age}} = 27.94$, $SD = 7.95$) and 523 males (60.25%; $m_{\mathrm{age}} = 20.09$, $SD = 9.29$). These groups did not differ according to age, ($t[868] = 1.86$, $p = 0.062$). Based on the binary recoding approach, 13 individuals (1.5%) met the diagnostic threshold for IGD (i.e., endorsing "very often" on five or more criteria).

Measure

Internet Gaming Disorder Scale 9-Short-Form As mentioned in the introduction, the Internet Gaming Disorder Scale 9-Short Form (IGDS9-SF) (Pontes and Griffiths 2015) is a short psychometric tool based on the nine core criteria defining IGD in the DSM-5 (American Psychiatric Association, 2013). The IGDS9-SF was originally developed for assessing the severity of online and/or offline gaming activities occurring over a 12-month period. Each item is answered using a five-point scale: 1 ("never"), 2 ("rarely"), 3 ("sometimes"), 4 ("often"), and 5 ("very often"). Pontes and Griffiths (2016) have suggested that for clinical diagnosis, the endorsement of five or more of the nine criteria in the IGDS9-SF, on the basis of answering "very often," should be considered, as this approach mirrors the diagnostic framework suggested by the APA in the DSM-5 (American Psychiatric Association, 2013). Given the goals of the current study, and consistent with the more stringent binary scoring method recommended by Pontes and Griffiths (2015), the five response options of the items were recoded so that for each item the first four response options (i.e., options 1 to 4) were deemed as absence of the criterion (i.e., no endorsement of the IGD criterion), while the next response option category (i.e., option 5 or "very often") was recoded as the criterion being present (i.e., endorsement of the IGD criterion) (Pontes and Griffiths 2015; Pontes et al. 2016). Internal reliability for the recoded IGDS9-SF for the sample in the present study was found to be adequate (Cronbach's alpha = 0.91).

Procedure

The study was approved by the Human Research Ethics Committee of the relevant institutions and participants were recruited online. Eligible individuals (residents or citizens of the USA) interested in participating were invited to register with the study via a *SurveyMonkey* link that was advertised across numerous online gaming forums ($n = 405$) and the crowdsourcing platform *Amazon Mechanical Turk* (AMT) ($n = 463$).

The link of the study directed potential participants to the plain language information statement (PLIS). The PLIS explicitly indicated that participation was voluntary and that participants were free to withdraw from the study at any time prior to its completion. Any discontinuation of participation, at any point, required no explanation and was without any penalties. Furthermore, participation in the study was only possible after participants had provided their consent and indicated that they fully understood the nature of the research being conducted. It is worth noting that online data collection was preferred over more traditional paper-and-pencil data collection based on relevant literature recommendations indicating that this method is cost-effective and facilitates accessibility to hard-to-reach groups (i.e., gamers) that were relevant to the present study (Griffiths 2010). Overall, research has shown that online data collection and paper-and-pencil methods are generally equally valid and reliable (Pettit 2002; Weigold et al. 2013).

Statistical Analysis and Data Analytic Strategy

The unidimensional (one-factor) model for the IGDS9-SF (an assumption in the 2PLM) was examined using the confirmatory factor analysis (CFA) procedure for ordered-categorical data, as recommended by Reeves et al. (2007) and Hill et al. (2004). Support for unidimensionality is inferred when there is good model fit, which is demonstrated by significant and substantial factor loadings. A weighted least squares estimator (WLSMVχ^2) was used to estimate all the CFA models through M*plus* version 7.11 (Muthen and Muthen 1998–2013). Like all other χ^2 values, WLSMVχ^2 values are prone to inflation when estimated with large sample sizes (see Bentler 1990). Therefore, other fit indices were used to determine model fit, including the root mean squared error of approximation (RMSEA) and the comparative fit index (CFI). Accordingly, model fit was based on guidelines suggested by Hu and Bentler (1999), with RMSEA values of 0.06 or below as a good fit, values > 0.06 to 0.08 as indicative of moderate fit, values > 0.08 to 0.10 as a marginal fit, and values > 0.10 indicative of a poor fit. For the CFI, values of 0.95 or above indicate good fit, values of > 0.90 and < 0.95 are acceptable fit, and values less than 0.90 are poor fit. Furthermore, statistical differences between the models assessed were tested using the difference in WLSMVχ^2 values. This study used the option available in *Mplus* to compute the WLSMVχ^2 difference values and the corresponding differences in the *df* values. An α value of .01 was used to allow for more stringent type 2 error control in the models compared.

In that line, the present study used item response theory for patient-reported outcomes (IRTPRO) Version 3.1 (Cai et al. 2011) to perform the 2PLM analyses. For each IGD item, the following IRT psychometric parameters were examined: ICC (graphically), α, β, and IIF (graphically). In addition, for the overall IGDS9-SF, the TIF was also examined (graphically). For the ICC, IIF, and TIF graphs, the *x* axis is the IGD trait scale from -3.00 to 3.00 *SD*, with mean = 0 and $SD = 1$.

As 2PLM is model-based, it is necessary to test if there is model-data fit. Item level fit to model was examined using the S-χ2 item-fit statistic for each item provided by IRTPRO. For each response category, S-χ^2 indicates the degree of similarity between model-predicted and empirical (observed) response frequencies, with a statistically significant value indicating poor fit. Given that this value is sensitive to large sample size, statistical significant cut-off value was set at $p = .01$ (Stone and Zhang 2003). In addition to item fit, the fit for the overall model was examined using the M_2 limited information goodness-of-fit statistic and its associated *p* value, and the RMSEA value (Cai et al. 2006). These values are also provided by IRTPRO upon request. M_2 values assume perfect model-data fit in the population, with smaller values indicating better fit. However, M_2 values are also sensitive to large sample sizes. Consequently, this study used *p* at .01 to infer statistical significance. The RMSEA values reported in IRTPRO are interpreted as in CFA models, with values close to .06 or less indicating good fit.

In addition to unidimensionality and model data fit, the 2PLM assumes local independence. Local independence implies that associations between items are only caused by the underlying latent trait. Local independence was examined using the standardized LDχ^2 statistic for each item pair provided by IRTPRO. Generally, LDχ^2 statistics greater than 10 are considered large and reflecting likely local dependence (Cai et al. 2011). In addition to the standardized LDχ^2 statistic, the output for the one-factor CFA model was also used. For this, support for local independence can be inferred when no residual correlations are more than .20 (Morizot et al. 2007), and when none of the modification indices (MI) for error covariances are abnormally large compared to the others (Hill et al., 2004) (Table 1).

Table 1 Problem players by item and per IGDS9-SF in total according to gender

		Female	Male	Total
Item 1	Problem behavior not present	324 39.5%	497 60.5%	821 100.0%
	Problem behavior present	21 44.7%	26 55.3%	47 100.0%
Item 2	Problem behavior not present	333 39.8%	503 60.2%	836 100.0%
	Problem behavior present	12 37.5%	20 62.5%	32 100.0%
Item 3	Problem behavior not present	329 39.5%	504 60.5%	833 100.0%
	Problem behavior present	16 45.7%	19 54.3%	35 100.0%
Item 4	Problem behavior not present	327 39.1%	510 60.9%	837 100.0%
	Problem behavior present	18 58.1%	13 41.9%	31 100.0%
Item 5	Problem behavior not present	332 39.6%	506 60.4%	838 100.0%
	Problem behavior present	13 43.3%	17 56.7%	30 100.0%
Item 6	Problem behavior not present	333 39.5%	510 60.5%	843 100.0%
	Problem behavior present	12 48.0%	13 52.0%	25 100.0%
Item 7	Problem behavior not present	338 39.9%	509 60.1%	847 100.0%
	Problem behavior present	7 33.3%	14 66.7%	21 100.0%
Item 8	Problem behavior not present	318 39.6%	486 60.4%	804 100.0%
	Problem behavior present	27 42.2%	37 57.8%	64 100.0%
Item 9	Problem behavior not present	339 40.1%	507 59.9%	846 100.0%
	Problem behavior present	6 27.3%	16 72.7%	22 100.0%
IGDS9-SF All Items	Problematic gaming not present	339 39.6%	516 60.4%	855 100.0%
	Problematic gaming present	6 46.2%	7 53.8%	13 100.0%

Responses equal to a score of 5 (“very often”) are reported as “problem behavior present” per item. More than five problem behaviors present per participant across all nine items are reported as “problematic gaming present”

All items were significantly and positively correlated with each ($p < .001$). All item loadings on the IGD latent factor were significant ($p < .001$)

Results

One-Factor Model (Unidimensionality) for the IGDS9-SF

Table 2 presents the tetrachoric correlation matrix between the nine IGDS9-SF criteria used in the CFA. As shown in the table, all the criteria were significantly and positively correlated ($p < .001$).

Table 2 Inter-tetrachoric correlations, descriptives, and items loadings for the nine IGD items

	IGD1	IGD2	IGD3	IGD4	IGD5	IGD6	IGD7	IGD8	IGD9
IGD1	–								
IGD2	.74	–							
IGD3	.56	.69	–						
IGD4	.66	.65	.79	–					
IGD5	.73	.73	.74	.77	–				
IGD6	.69	.70	.64	.67	.72	–			
IGD7	.53	.61	.64	.75	.72	.67	–		
IGD8	.57	.61	.47	.38	.56	.45	.45	–	
IGD9	.56	.77	.46	.49	.66	.60	.78	.66	–
Mean (SD)	.05 (.226)	.04 (.189)	.04 (.197)	.04 (.186)	.03 (.183)	.03 (.167)	.02 (.154)	.07 (.261)	.03 (.157)
Factor loading	.80	.87	.81	.85	.88	.80	.82	.65	.82

All items were significantly and positively correlated with each ($p < .001$). All item loadings on the IGD latent factor were significant ($p < .001$)

The fit values for the one-factor measurement model with all the nine IGDS9-SF criteria indicated good fit (WLSMVχ^2 [$df = 27$] = 41.67, $p = .035$; RMSEA = .025 [90% confidence interval = 0.007, 0.039 and CFI = .986) (Hu and Bentler, 1999). It is notable that the chi-square estimator (WLSMVχ^2) was not significant at $p = .01$, indicating that the data did not depart from the specified model under the $p < .01$ significance level. Additionally, none of the MI values between the error variances was extremely large. Specifically, the largest MI (i.e., 6.54) was between criteria 4 and 3; the fit values of a revised measurement model with the error variances for these criteria correlated were WLSMVχ^2 ($df = 26$) = 35.96, $p = .092$; RMSEA = .021, 90% CI [0.000, 0.036]; CFI = .991, and the WLSMVχ^2 value for this model did not differ significantly, at $p = .01$, from the initial model without the addition of the correlation between the two error terms aforementioned (ΔWLSMVχ^2 [$\Delta df = 1$] = 4.407, $p = .36$). Table 2 also shows the standardized factor loadings of all the nine criteria on the IGD latent factors. As shown, all loadings were high (ranging from .65 to.88) and statistically significant ($p < .001$). The findings obtained provide compelling support for the unidimensionality of the IGDS9-SF with all the nine criteria.

Evaluation of Local Independence and Model-Data Fit for the IGDS9-SF Items

Prior to the examination of the IRT parameters, support for local independence and model data fit were examined for this model.

Local Independence In relation to local independence, the standardized LDχ^2 statistic for each criterion pair provided by IRTPRO ranged from 0.0 to 1.0. Additionally, the CFA of this model with the nine criteria indicated that the highest residual correlation was .19, and the remaining residual correlations ranged from .01 to .16. In addition, as noted above, for the initial CFA model with the nine criteria, the MI between the error variances were all relatively small (i.e., < 6.54). These findings support the assumption of local independence for all nine criteria modeled in the 2PLM.

Model Data Fit Table 3 shows the S-χ^2 item-fit statistic for each criterion provided by IRTPRO from the calibration results. As shown, the S-χ^2 statistic for all nine criteria indicated satisfactory fit (i.e., all *p*s were > .023). In relation to fit at the overall model level, the M_2 value was not significant at $p = .01$ (M_2 [27] = 45.12, $p = .016$), and the RMSEA value was within the expected range for a well-fitting model (0.03). Taken together, these findings provide evidence of good model data fit. Therefore, in addition to supporting the scale's unidimensionality, there is also support for local independence of the nine criteria, and model data fit for the 2PLM tested.

Item Parameter Estimates for the IGDS9-SF Items

The discrimination and difficulty parameters for the nine criteria in the 2PLM are provided in Table 3. Figure 2 shows the ICCs and IIF for these criteria.

Discrimination Parameters Although there was wide variability in strength of discrimination across IGDS-9SF items, the discrimination values for all the criteria were high, ranging from 1.57 to 3.57. These values indicate that each item demonstrated a good capacity when discriminating the underlying IGD dimension (Table 3). The high α values and the differences for these values are illustrated in Fig. 1 by the steep slopes for the ICCs, and the variations in the slopes of these curves across all the nine IGD criteria. The order of the criteria in terms of increasing discrimination values were IGDS9-SF items 8, 1, 6, 9, 3, 4, 7, 2, and 5, with values of 1.57, 2.41, 2.47, 2.50, 2.57, 2.75, 2.82, 3.29, and 3.57, respectively. Thus, compared to the other criteria, criterion "escape" (item 8) had noticeable lower α value, while criteria "giving up other activities" (item 5) and "withdrawal" (item 2) had relatively higher α values.

Table 3 and Fig. 2 demonstrate that although there was some degree of variability for the difficulty values, all items were located from + 2.00 to + 2.41 *SD* from the mean. The order of the items in terms of increasing difficulty values were IGDS9-SF items 1, 2, 5, 3, 4, 8, 7, 6, and 9, with values of 2.00, 2.03, 2.03, 2.13, 2.15, 2.18, 2.34, 2.35, and 2.41 *SD*s from the mean, respectively. The difficulty values determine the point at which the response curves intersect on the θ-scale. As an illustration for interpreting these graphs, for item 1, for a θ score < + .2.00, there is a 50% probability that the "no" response is more likely to be endorsed by the individual (i.e., absence of the IGD criterion), and for a θ score > + .2.00, there is a 50%

Table 3 Two-Parameter Logistic Item Response Model Parameter Estimates and S-χ^2 Statistics

	IRT Estimates				S-χ^2 Statistics		
Item (brief description)	*a*	*SE*	*b*	*SE*	S-χ^2	*df*	*p*
IGD1 (preoccupation)	2.41	0.46	2.00	0.20	0.84	5	0.9743
IGD2 (withdrawal symptoms)	3.29	1.05	2.03	0.20	2.95	4	0.5668
IGD3 (tolerance)	2.57	0.53	2.13	0.20	4.72	4	0.3180
IGD4 (loss of control)	2.75	0.64	2.15	0.21	4.18	5	0.5254
IGD5 (giving up other activities)	3.57	0.78	2.03	0.17	2.36	4	0.6701
IGD6 (continuation)	2.47	0.59	2.35	0.24	3.26	5	0.6610
IGD7 (deception)	2.82	0.59	2.34	0.22	14.58	6	0.0238
IGD8 (escape)	1.57	0.25	2.18	0.23	3.19	4	0.5280
IGD9 (negative consequences)	2.50	0.55	2.41	0.22	1.67	5	0.8933

Note. α = discrimination parameter; β, = difficulty parameter

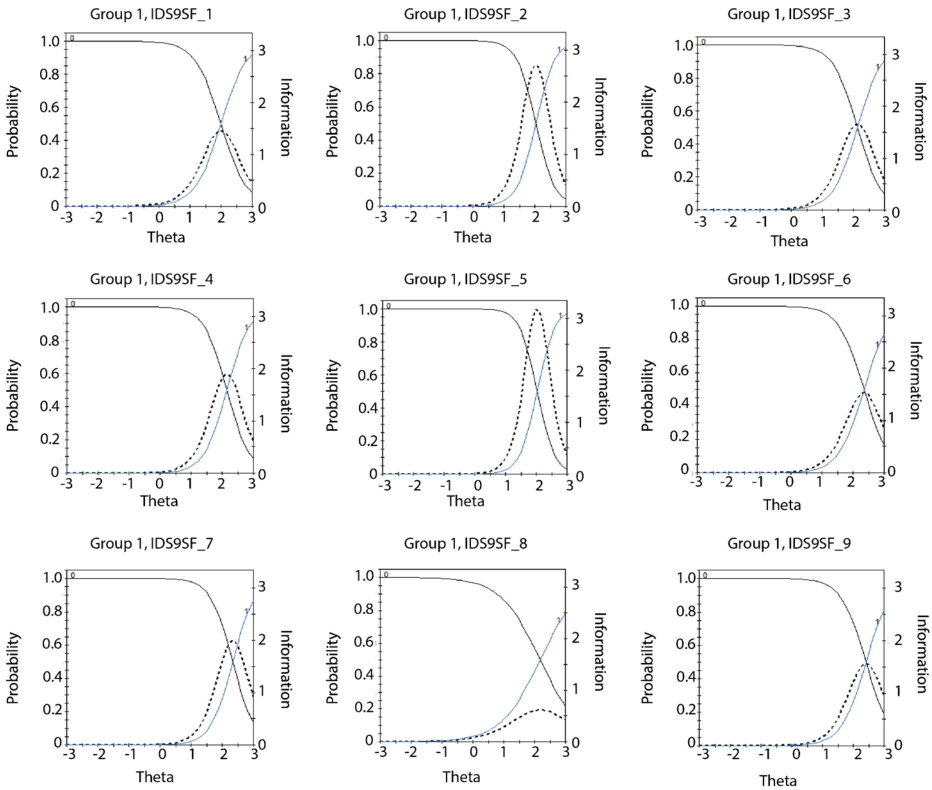

Fig. 2 Item characteristic curve (ICC indicate as curve 1 in figure) and item information function (dotted) of the IGDS9-SF symptoms

probability that the "yes" response (i.e., presence of IGD criterion) is more likely to be endorsed by the individual. Thus, compared to the other criteria, criteria "preoccupation" (item 1), followed by "withdrawal" (item 2) and "giving up other activities" (item 5) were endorsed at relatively lower IGD trait levels, since they had lower difficulty values. In contrast, criteria "continuation despite problem" (item 6), "deception" (item 7), and "negative consequences" (item 9) required relatively higher levels of the IGD trait to be endorsed as they had relatively higher difficulty values.

Item Information Function Figure 2 shows that IIF values for all criteria were around 0 up to about + 0.5 *SD* from the mean. All criteria except criterion "*escape*" (item 8) demonstrated a reasonable amount of item information (and therefore reliability) from around the + 1.5 to around + 3 *SD* from the mean, peaking at around + 2 *SD* to + 2.5 *SD* from the mean. From around the + 1.5 to around + 3 *SD* from the mean, criteria "withdrawal" (item 2) and "giving up other activities" (item 5) had relatively high information values (and therefore reliability). Further results indicated that criteria "preoccupation" (item 1), "tolerance" (item 3), "loss of control" (item 4), "continuation despite problems" (item 6), "deception" (item 7), and "negative consequences" (item 9) had moderately high information values. Finally, criterion “escape” (item 8) had relative lower information values (and therefore reliability), compared to other IGD criteria across the entire trait spectrum.

Test Characteristic Curve and Test Information Function for the IGDS9-SF

Broad-based TCC and TIF graphs for the IGDS9-SF were conducted to illustrate criterion endorsement across the measure (see Fig. 3). The TCC graph illustrates that the trait of IGD increased monotonically steeply, as the number of criteria endorsed increased from 1 to 9. The steep curve suggests that the IGDS9-SF is an adequate psychometric tool for identifying individuals with high and low levels of the IGD trait. The IGD trait level for endorsement of at least five criteria suggests that an IGDS9-SF total score + 2 *SD* from the mean on the IGD trait spectrum would be likely to identify individuals who qualify for the IGD diagnosis. Consistent with the IIF values, Fig. 3 shows that for the IGDS9-SF scale, as a whole, the TIF values were around 0 up to about + 0.5 *SD* from the mean, and that there was reasonable amount of test information from around + 1.5 to around + 3 *SD* from the mean, peaking at around + 2 *SD* from the mean.

Discussion

The aim of the study was to examine the psychometric properties of the nine IGD criteria as measured by the IGDS9-SF in a large and heterogeneous sample of adolescent and adult USA gamers using IRT analysis (2PLM). As for the IRT analysis (based on the binary recoding process outlined for the IGDS9-SF), the results indicated that the IGDS9-SF items designed to assess IGD criteria are psychometrically sound, as they demonstrated excellent discrimination values. For most criteria measured by the IGDS9-SF, difficulty values were high (at above + 2 *SD* from the mean). The IGD criteria “giving up other activities” (item 5) and “withdrawal” (item 2) had especially higher discrimination values ($\alpha = + 3.57$ and $+ 3.29$, respectively), thereby indicating that they were stronger items when used to discriminate individuals with and without high levels of the IGD trait. These results converge with theoretical assumptions

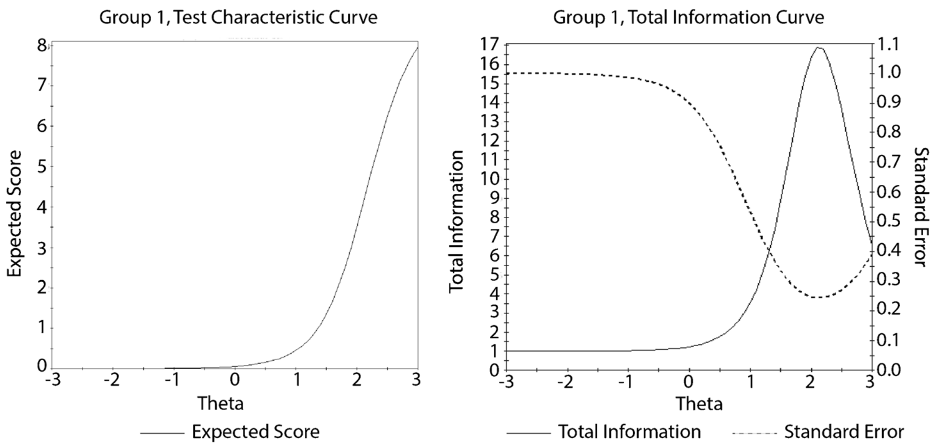

Fig. 3 Test characteristic curve (left) and test information function and its standard error (dotted line; left) for the IGDS9-SF

about behavioral addictions and recent empirical findings. More specifically, these finding support Griffiths' (2005) components model of addiction [which has been demonstrated to overlap with the nine IGD criteria theoretically (see Pontes and Griffiths) and empirically (see Pontes et al. 2014)] that argues that "conflict" (IGD criterion 5) and "withdrawal" (IGD criterion 2) are core features of behavioral addictions such as IGD. Additionally, these findings lend empirical support to recent psychophysiological evidence supporting the presence of withdrawal-like effects in behavioral addictions (see Reed et al. 2017).

The results also revealed that for most IGD criteria as measured by the IGDS9-SF, information values were extremely low up to around + 0.5 *SD* from the mean relative to their reliabilities at + 2 *SD* to + 2.5 *SD* from the mean. Furthermore, the overall findings for the discrimination values in this study suggest that the nine IGD criteria, as measured by the IGDS9-SF, are able to effectively discriminate those with high and low levels of IGD trait. The findings for the difficulty values suggest that for these recoded scores, there is a 50% probability that the criteria would be endorsed as being present when their underlying trait levels are at least around + 2 to + 2.5 *SD* from the mean with adequate reliabilities, at around the mean to + 2 *SD* from the mean. It is important however to note that these are reliabilities of the different IGDS9-SF criteria to measure the IGD and not their own individual unique reliabilities (e.g., the reliability of an item measuring its referred criterion). Notwithstanding this, this finding is of utmost clinical importance and underscores the advantages of applying IRT to the investigation of IGD criteria as psychometric standardized tools designed to assess this phenomenon should be able to reliably discriminate at different levels of the disorder, something that is not tenable to be examined using CTT-based approaches.

Despite a reasonable level of consistency in the results for the item parameters, there were key differences between the parameter estimates for the criteria worthy of specific note. The results showed that although the discrimination (α) values across the different criteria were high, they varied noticeably (ranging from 1.57 to 3.57), thereby indicating the IGD criteria as measured by the IGDS9-SF differed in their ability at discriminating those gamers with low and high levels of the IGD trait. The order of the criteria in terms of increasing discrimination values were IGDS9-SF items 8, 1, 6, 9, 3, 4, 7, 2, and 5. The values were 1.57, 2.41, 2.47, 2.50, 2.57, 2.75, 2.82, 3.29, and 3.57, respectively. Based on guidelines proposed by Baker (2001) that discrimination values < 0.64 are considered low, and that discrimination values from 0.65 to 1.34 are moderate, from 1.35 to 1.69 are high, and > 1.69 are perfect (interpreted here as very high), all IGD criteria, except criterion "escape" (item 8), can be considered high. "Escape" had noticeable lower ability to discriminate those with and without high levels of the IGD trait. In contrast, criteria "giving up other activities" (item 5) and "withdrawal" (item 2) had relatively higher discrimination values ($\alpha = 3.57$ and 3.29, respectively), and therefore, higher ability to discriminate those with and without high levels of the IGD trait. There was also some degree of variability for the difficulty values, with values ranging from around + 2 SD to + 2.41 SD from the mean. However, as the differences in the difficulty parameter values were all within 0.5 SD, the differences could be taken as small. Thus, the findings indicated that generally for the recoded IGDS9-SF criteria, there is a 50% probability that they would be endorsed as being present, when their underlying IGD latent trait levels are around + 2 SD or slightly more than + 2 SD from the mean, depending on the criterion. Additionally, unlike the other IGD criteria, criterion "escape" (item 8) had low reliability for virtually all levels on the trait spectrum.

Similarly to the findings obtained in the current study, the study by Wu et al. (2017a) that applied the Rasch model on the ratings of the IGDS9-SF, as well as the study by Király et al.

(2015) that examined the 2PLM properties for the recoded scores of the IGDT-10 (Király et al. 2015) have reported wide variability considering the item difficulty and/or discrimination parameters. However, there were important similarities and differences between the findings of these studies in comparison to the present study. First, relative to most of the other criteria, Király et al. (2015) found a lower discrimination parameter value for criterion “escape” (item 8), and a higher discrimination parameter value for item 2 (“withdrawal”). It was also found that relative to most of the other IGD criteria, criterion “preoccupation” (item 1) had a lower difficulty parameter value. Third, in line with the findings reported by Király et al. (2015), we also found relative low difficulty parameter value for criterion “escape” (item 8).

Although, findings for the parameter estimates of the current study were directly compared with the findings in previous studies, caution is advised when interpreting the present results, as IRT utilizes responses to individual items to obtain continuous scaled estimates of the underlying trait or theta. In most statistical modeling packages, the trait value is set at a mean of zero and a standard deviation of one. Since the difference in the response categories and scoring methods across the measures used in different studies could lead to different responses by the same individual, the metrics for the measures would be different. Consequently, their parameters cannot be directly compared. Thus, all IRT findings reported in this study are novel, and extend existing psychometric information for the IGDS9-SF that has been derived essentially via CTT-based studies.

Clinical and Diagnostic Implications

The findings obtained in the present study may have key clinical and diagnostic implications for the assessment IGD and potential implications for revising (or at least investigating further) the nine IGD diagnostic criteria. More specifically, criterion “escape” (item 8) had relatively low discrimination ability, and its reliability was relatively low at all trait levels. The latter means that this criterion does not measure IGD trait with sufficient accuracy, and may need some revision or even removal from future revisions of the IGD diagnostic criteria. While this study suggests the need to revise the criterion for “escape,” suggesting exactly how it should be revised is beyond the scope of the present study as a clinical sample would be required. Notwithstanding this, at a more general level, the low reliability value for the criterion “escape” may also have direct implications for the use of this criterion in the diagnosis of IGD as it may not be reliable for the diagnosing IGD. Our findings suggest that its adequacy as a diagnostic criterion may need to be reviewed in future editions of the DSM. Despite the preliminary evidence found here with regards to the inadequacy of the “escape” criterion, this finding needs to be interpreted with caution as further clinical research is necessary to corroborate or invalidate this finding as the present sample was a community-based sample. Additionally, as noted by Wender (2004), unlike rating scales, clinical interviews provide opportunities for clinicians to deal with respondents’ uncertainties when answering questions.

The present findings with regards to the criterion “escape” echo those reported by Ko and colleagues (2014) using a clinical sample of disordered gamers ($n = 75$), remitted disordered gamers ($n = 75$), and a control group ($n = 75$) that found that the criterion “escape” (alongside “deception”) presented with the poorest diagnostic accuracy values in comparison to all other criteria. A potential explanation for this finding could be related to the fact that “escape” may be best understood as a gaming motivation and risk factor for disordered gaming (as opposed to a core criterion of IGD), further implying that although “escape” maybe a reliable predictor of IGD as found by several studies (e.g., Király et al. 2015; Wu et al. 2017a), it may be better understood as a peripheral feature (i.e., underlying motivation) and not a central aspect of IGD.

The results of this study also have implications for the future use of the IGDS9-SF in research and clinical settings. Given the IRT results, it can be argued that when IGDS9-SF criteria are used as binary scores, they would generally provide a highly discriminative and reliable measure of their underlying IGD latent trait for those with high levels of IGD. In this respect, as item 5 (criterion "giving up other activities") and item 2 (criterion "withdrawal") demonstrated sound ability to discriminate those with and without high levels of the IGD trait, and reliable representation of the IGD trait at high levels, these criteria may be considered to be important for the identification and diagnosis of IGD. This is an important finding that corroborates recent studies using different methodological approaches with regards to the clinical utility of the criteria "giving up other activities" and "withdrawal." More specifically, the study by Rehbein et al. (2015) using a large representative sample of 11,003 ninth-graders aged between 13 to18 years from Germany found that "give up other activities" best corresponded with the full IGD diagnosis and that this criterion alongside "withdrawal" were the most relevant and useful criteria for IGD diagnosis. Additionally, Rehbein et al. (2015) also found that although "escape" was endorsed most frequently by their sample, it rarely related to IGD diagnosis, further supporting the present findings with regards to this criterion.

Further implications of the IRT analysis suggest that overall the nine IGD criteria outlined in the DSM-5 are captured in the IGDS9-SF, and more information regarding the validity of the criteria used for diagnosis can be derived from prior and future use of the instrument. Furthermore, IGD can be assessed from a dimensional and categorical vantage using this instrument. The results obtained here may pave the way to future research as the present findings may be utilized to discern important screening items for brief screening an intervention of IGD in clinical prevention and intervention research.

Additional implications may be related to the fact that a general rule of thumb in psychological measurement is that scores 2 *SD* from the mean in the deviant direction are considered clinically meaningful. Since the difficulty parameter values of all the criteria in the IGDS9-SF were at this point or higher, it can be substantiated that an IGD score of + 2 *SD* may be able to efficiently distinguish those with and without high levels of IGD, which is of utmost importance to any clinical assessment tool.

In addition to this, it is worth noting that the findings that the IGD criteria are generally unreliable and do not adequately represent the appropriate trait, from close to the mean to low trait levels, implies that measuring IGD criteria with the IGDS9-SF may result in unreliable scores for individuals with relatively low levels of the IGD criteria. Thus, its use with individuals with low levels of the IGD criteria may be problematic. This may be particularly relevant for community-based studies and particularly advantageous to clinical-based studies screening individuals with potentially elevated IGD-related symptomatology. Despite this, the use of such measures can still be considered appropriate for epidemiological and prevalence studies since the focus of such studies is not on individuals with low levels of the IGD criteria, but on individuals with high levels of these criteria. The primary goal of such studies is to ascertain prevalence rates of specific disorders in the broader population.

An approach that has been proposed for scoring the IGDS9-SF is to use the total scale scores (Pontes and Griffiths 2015). Given that there was some degree of variability in the difficulty parameters, some may consider that the algebraic summation of unweighted raw scores of the criteria to obtain the total score is mathematically inappropriate. Put simply, assigning the same clinical weight to all nine criteria may be inappropriate given the findings encountered in the present study. Nevertheless, it is argued here that as the variability in the difficulty parameters across the criteria was small, and as the relationships between IRT estimated theta values and the total scores were (fairly) linear from + 2 *SD* onwards (see Fig. 3), this may not be a problem from a practical viewpoint. However, the findings presented here also showed that IGDS9-SF criteria are generally not reliable and are weak at representing the appropriate traits at low trait levels. As the total score is based on all criteria in the IGDS9-SF, it follows that total scores will include criteria with endorsements of lower responses options. Given that at this level there is low reliability and weak representation of the appropriate traits, it can be argued that total scores have questionable utility as they assume all criteria measured by the nine criteria are equally relevant and important toward IGD diagnosis. Thus, the use of total scores may not be a useful approach for inferring whether an individual would potentially qualify for the IGD. It is therefore suggested that when using the IGDS9-SF, the binary recoded scores (as suggested by Pontes and Griffiths 2015 and applied in the present study) may be a best procedure to be used.

In summary, the major contribution of the study is that this is one of the few studies to provide IRT properties of the IGD criteria as measured by the IGDS9-SF. The IRT findings indicated that all the criteria assessed by the psychometric test were strong discriminators of the IGD trait. Additionally, they measured more of the IGD trait variance and with more precision in the upper half of the trait continuum, which is useful for clinical and epidemiological studies. Despite these new findings, there are also potential limitations in the study that need to be acknowledged. First, as this study examined gamers from online communities, the findings may not be totally applicable to clinical samples. Second, it is important to keep in mind that the information for the IGDS9-SF were derived from self-ratings, which may be affected by common method. However, modeling common method variance effects in IRT procedures is complex, and was not possible in the 2PLM used in this study. Thus, it is not certain how method-related effects could have confounded the results in this study. Third, to the extent that this study examined the recoded scores of the IGDS9-SF in a USA sample, the findings must be seen as limited to gamers from this particular country. Additionally, the relevance of the findings and conclusions made here for the IGDS9-SF to other IGD rating scales or interview-based data in other national groups are uncertain and warrant further research. Fourth, as the IGDS9-SF is a clinical measure to facilitate the diagnosis of IGD, it would be useful to replicate this study with individuals diagnosed or with high levels of IGD criteria. Finally, the appropriateness of the application of 2PLM in the study could be questioned. This model assumes that traits are bipolar, that is, both ends of the trait continuum scale represent meaningful variations of the trait. Thus, the mean of the latent trait is defined as zero, with low scores reflecting levels below the average levels. According to Reise and Waller (2009), many clinical constructs could be unipolar, where one end of the trait continuum represents severity and the other end represents its absence. Lucke (2014) has suggested that for such traits, the person with a certain amount of the trait has to be referenced to the level of no trait, and not the mean. This means that low scores represent the absence of the trait and not scores below the average, and thus zero is the lowest possible latent trait score. Interestingly, he developed new

IRT models (called unipolar item response models) and illustrated their applications with reference to a gambling addiction scale. Although such models may seem as viable alternative to the 2PLM for application in the current study, Lucke (2014) has pointed out that the assumption in unipolar item response models, that the probability of item endorsement is zero for those individuals with a trait level at zero, does not necessarily apply to other unipolar traits. Thus, it does not make sense to diminish the relevance of the 2PLM for the current study. Given these limitations, there is a need for additional cross-validation of the findings, keeping in mind the potential limitations discussed here. Despite these potential concerns, at the more general level, this study has shown that the use of IRT procedures can provide valuable additional psychometric information, and also inform practical and theoretical issues relevant for IGD and for clinical psychology in general. It is envisaged by the authors of the study that these findings will facilitate future research using IRT-based models for evaluating the psychometric properties of the different IGD measures that are now available.

Authors' Contribution RG has contributed to the statistical analysis, literature review, conceptualization of the study, and writting up of the manuscript. VS has contributed to the literature review, data collection and statistical analysis, structure and sequences of theoretical arguments and formatting. CB has contributed to the data collection, theoretical argument, and manuscript review. HMP has contributed to the literature review, hypotheses formulation and revising the manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Standards and Animal Rights All procedures performed in the study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants.

Confirmation Statement Authors confirm that this paper has not been either previously published or submitted simultaneously for publication elsewhere.

References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders *(5th ed.).* Arlington, VA: American Psychiatric Publishing.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.
Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173–194. https://doi.org/10.1348/000711005X66419.
Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible professional item response theory modeling for patient reported outcomes (Version 2.1) [Computer software]*. Chicago: Scientific Software International.
Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.
Griffiths, M. (2005). A 'components' model of addiction within a biopsychosocial framework. *Journal of Substance Use, 10*(4), 191–197. https://doi.org/10.1080/14659890500114359.
Griffiths, M. D. (2010). The use of online methodologies in data collection for gambling and gaming addictions. *International Journal of Mental Health and Addiction, 8*, 8–20. https://doi.org/10.1007/s11469-009-9209-1.
Hawi, N. S., & Samaha, M. (2017). Validation of the Arabic version of the internet gaming disorder-20 test. *Cyberpsychology, Behavior, and Social Networking, 20*(4), 268–272.
Hill, C. D, Edwards, M. C, Thissen D, Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2004). Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 Generic Core Scales. *Medical Care, 45* (Suppl 1), S39 –S47. doi: https://doi.org/10.1097/01.mlr.0000259879.05499.eb.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.
Keeley, J. W., Reed, G. M., Roberts, M. C., Evans, S. C., Medina-Mora, M. E., Robles, R., & Andrews, H. F. (2016). Developing a science of clinical utility in diagnostic classification systems: field study strategies for ICD-11 mental and behavioral disorders. *American Psychologist, 71*(1), 3–16. https://doi.org/10.1037/a0039972.
Király, O., Urbán, R., Griffiths, M. D., Ágoston, C., Nagygyörgy, K., Kökönyei, G., & Demetrovics, Z. (2015). The mediating effect of gaming motivation between psychiatric symptoms and problematic online gaming: an online survey. *Journal of Medical Internet Research, 17*(4), e88.
Lucke, J. F. (2014). 13 unipolar item response models. Handbook of Item Response Theory Modeling, 272.
Monacis, L., Palo, V. D., Griffiths, M. D., & Sinatra, M. (2016). Validation of the internet gaming disorder scale–short-form (IGDS9-SF) in an Italian-speaking sample. *Journal of Behavioral Addictions, 5*(4), 683–690. https://doi.org/10.1556/2006.5.2016.083.
Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality* (pp. 407–423). New York: Guilford Press.
Muthén, L. K., & Muthén, B. O. (2013). Mplus 7.11. Los Angeles, CA*: Muthén & Muthén*.
Pettit, F. A. (2002). A comparison of world-wide web and paper-and-pencil personality questionnaires. *Behavior Research Methods, Instruments, & Computers, 34*(1), 50–54. https://doi.org/10.3758/BF03195423.
Pontes, H. M. (2016). Current practices in the clinical and psychometric assessment of internet gaming disorder in the era of the DSM-5: a mini review of existing assessment tools. *Mental Health and Addiction Research, 1*(1), 18–19. https://doi.org/10.15761/MHAR.1000105.
Pontes, H. M., & Griffiths, M. D. (2014). Assessment of internet gaming disorder in clinical research: past and present perspectives. *Clinical Research and Regulatory Affairs, 31*(2–4), 35–48. https://doi.org/10.3109/10601333.2014.962748.
Pontes, H. M., & Griffiths, M. D. (2015). Measuring DSM-5 internet gaming disorder: development and validation of a short psychometric scale. *Computers in Human Behavior, 45*, 137–143. https://doi.org/10.1016/j.chb.2014.12.006.
Pontes, H. M., & Griffiths, M. D. (2016). Portuguese validation of the internet gaming disorder scale–short-form. *Cyberpsychology, Behavior, and Social Networking, 19*(4), 288–293. https://doi.org/10.1089/cyber.2015.0605.
Pontes, H. M., Király, O., Demetrovics, Z., & Griffiths, M. D. (2014). The conceptualisation and measurement of DSM-5 internet gaming disorder: the development of the IGD-20 test. *PLoS One, 9*(10), e110137. https://doi.org/10.1371/journal.pone.0110137.
Pontes, H. M., Macur, M., & Griffiths, M. D. (2016). OR-85: construct validity and preliminary psychometric properties of the internet gaming disorder scale short-form (IGDS9-SF) among Slovenian youth: a nationally representative study. *Journal of Behavioral Addictions, 5*(S1), 35–36. https://doi.org/10.1556/2006.5.2016.042.

Pontes H., Stavropoulos V. & Griffiths M. (2017). Measurement invariance of the internet gaming disorder scale–short-form (IGDS9-SF) between the USA, India and the UK. Psychiatry Research, *in press.*

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* London, New York: Routledge.

Reed, P., Romano, M., Re, F., Roaro, A., Osborne, L. A., Viganò, C., et al. (2017). Differential physiological changes following internet exposure in higher and lower problematic internet users. *PloS one, 12*(5), e0178480.

Reeves, B. B., Hays, R. D., Bjorner, J. B., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information systems (PROMIS). *Medical Care, 45*(Suppl 1), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.

Rehbein, F., Kliem, S., Baier, D., Mößle, T., & Petry, N. M. (2015). Prevalence of internet gaming disorder in German adolescents: diagnostic contribution of the nine DSM-5 criteria in a state-wide representative sample. *Addiction, 110*(5), 842–851. https://doi.org/10.1111/add.12849.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553.

Sharp, C., Steinberg, L., Yaroslavsky, I., Hofmeyr, A., Dellis, A., Ross, D., & Kincaid, H. (2012). An item response theory analysis of the problem gambling severity index. *Assessment, 19*(2), 167–175. https://doi.org/10.1177/1073191111418296.

Stavropoulos V., Beard C., Griffiths M., Burleigh T., Gomez R. & Pontes H. (2017). Measurement invariance of the internet gaming disorder scale–short-form (IGDS9-SF) between Australia, the USA, and the UK. *International Journal of Mental Health and Addiction, in press.*

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske.* Hillside, NJ: Lawrence Erlbaum Associates.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*, 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x.

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of selfreport survey-based paper-and-pencil and internet data collection methods. *Psychological Methods, 18*(1), 53–70. https://doi.org/10.1037/a0031607.

Wender, E. (2004). ADHD symptoms and parent-teacher agreement. *Journal of Developmental & Behavioral Pediatrics, 25*(1), 48–49.

Wu, A. M. S., Lai, M. H. C., Yu, S., Lau, J. T. F., & Lei, M.-W. (2017a). Motives for online gaming questionnaire: its psychometric properties and correlation with internet gaming disorder symptoms among Chinese people. *Journal of Behavioral Addictions, 6*(1), 11–20.

Wu, T. Y., Lin, C. Y., Årestedt, K., Griffiths, M. D., Broström, A., & Pakpour, A. H. (2017b). Psychometric validation of the Persian nine-item internet gaming disorder scale–short form: does gender and hours spent online gaming affect the interpretations of item descriptions? *Journal of Behavioral Addictions, 6*(0), 1–8. https://doi.org/10.1556/2006.6.2017.025.